6P-6

人間認知の適応特性を実装した価値関数とその囲碁 AI への応用

西村友伸† 大用庫智† 高橋達二†

A loosely symmetric value function implementing adaptive cognitive properties of human and its application to Go AI

TOMONOBU NISHIMURA† KURATOMO OYO†
TATSUJI TAKAHASHI†

1. 序論

既存の研究において，可変参照型緩対称性推論（以下 LSVR ）をモンテカルロ木探索へ利用したリバーシの AI が，少ない試行回数で囲碁等の AI によく利用される UCT を利用したものより強い事が分かっている[7]．しかし，LSVR が既存のモデルと組み合わせた時の結果についてはまだ研究されていない．本研究では従来利用されてきた UCT と LSVR を組み合わせたハイブリッドモデルを作成しその効果を調べた．また LSVR は実際に囲碁のような広大な盤面を持つ競技に用いられたことはない．そのためリバーシ盤面を広げた場合について調査し，囲碁への応用が実用的である事を示す．

2. UCT

囲碁等のゲーム AI によく利用される UCT について説明する[3]. UCT は UCB1 [2] というスロットマシンの選択アルゴリズムを，木探索へ拡張したものである．UCB1 は全てのマシンを一度プレイし，その後は式（1）で計算された値が最も高いマシンをプレイしていく．X_i はマシン A_i の期待値，n_i はマシン A_i の選択回数，n は総選択回数を表している．c は第二項の重みを表し，c が大きいと選択回数の少ないマシンを選択する傾向が強まる．UCT では，各ノードから展開される子ノードをスロットマシンと見なし，UCB1 により最良のマシンであると計算された子ノードへ状態を遷移させる．この遷移を根ノードから葉ノードまで行い最良と考えられる葉ノードを得る．葉ノードでは何かしらの処理を実行し，その結果得た報酬を葉ノードとその全ての親ノードに反映させる．また各葉ノードの処理回数が一定の閾値を超えると木が成長し，そこから新たな子ノードが展開される．

$$UCB1(A_i) = X_i + c\sqrt{\frac{\log n}{n_i}} \qquad (1)$$

2.1 LSVR

人間認知の適応特性を利用した因果推論モデルとして LSVR が近年研究されている[4]．LSVR では表 2 を利用し式（6）で計算される最大の値を持つ選択肢を選ぶ．

表 2 LSVR で用いる分割表

	結果 W	結果¬W
選択肢 A_1	a_1	b_1
⋮	⋮	⋮
選択肢 A_n	a_n	b_n

$$S_p = \frac{b_{\max} b_{\min}}{b_{\max} + b_{\min}} \quad (4) \qquad S_n = \frac{a_{\max} a_{\min}}{a_{\max} + a_{\min}} \quad (5)$$

$$LS(W \mid A_i) = \frac{a_i + S_p}{a_i + b_i + \rho_R (S_p + S_n)} \qquad (6)$$

$$\rho_R = \frac{1}{R_t} - 1 \quad (7) \qquad R_{t+1} = \alpha R_t + (1-\alpha) r_t \quad (8)$$

ここで a_{max}，b_{max} は試行回数が最も多い選択肢の a_i と b_i を表し，a_{min} ，b_{min} はその逆である．R_t は時刻 t での LSVR の価値基準であるリファレンス報酬を意味し，時刻 t が 0 の時に初期値 0.5 の値をとる． R_t から算出される ρ_R は，実際に LSVR の判断の基準となる参照点のパラメータである． α はリファレンス報酬に対する学習率で 0.0 から 1.0 の実数とする． α が大きいほど現在の価値基準を重視し，その価値基準からの変化が小さくなる．r_t は時刻 t で得た報酬であり今回扱うリバーシであれば勝ちの時は 1 ，負けの時は 0 を受け取る事になる．LSVR はマシンの数が多い場合，少ない試行回数において UCB1 より良い結果を出している．木探索への拡張は UCT 同様であるが，遷移をする際に用いる値を LSVR で求めている．LSVR は UCB1 に比べ，早い段階では良い選択肢を選べる事，選択肢の数に結果が影響されにくい事が分かっている．

† 東京電機大学
Tokyo Denki University

Copyright ©2013 Information Processing Society of Japan.
All Rights Reserved.

3. 提案モデル：ハイブリッドモデル

既存の研究により，UCT と LSVR を実装した AI をリバーシ上で対戦させると，少ないプレイアウト数では LSVR が，多いプレイアウト数では UCT がそれぞれ優位に立つことが示されている[7]．そこで今回は各選択郡について試行回数が少ない時は LSVR ，多い時には UCT を使うハイブリッドモデルを提案する．

4. シミュレーション 1

UCT を用いたモデルと，ハイブリッドモデルをリバーシ上で対戦させる．一手あたりのサンプリングの回数は 2, 3, 5, 10, 15, 20, 30, 50, 100, 150, 300, 500, 1000, 5000, 10000, 50000 回とし，先手後手により有利不利が生じる為，先手後手入れ替え 100 局，計 200 局を対戦しハイブリッドモデルの勝率を調べた．UCB1 のパラメータ c を 0.1，LSVR の学習率 α を 0.8，成長の閾値を 1 回とし，ハイブリッドモデル内でのモデルの切り替え閾値を 50 とした．

4.1 シミュレーション 2

LSVR は選択肢が多い時，UCB1 と比べよりよい結果を出す事が分かっている．その為盤面の狭いリバーシより盤面の広い囲碁の様なゲームでより良い結果が期待出来る．そこで UCT と LSVR を盤面のサイズを 6x6, 10x10 と変更し先手後手入替え 300 局，計 600 局リバーシで対戦させ，囲碁への応用の妥当性を調べた．パラメータについては前節と同様とする．

5. シミュレーション 1: 結果

シミュレーション 1 の結果を図 1 に示す．縦軸はハイブリッドモデルの UCT に対する勝率を表し，横軸は一手あたりのプレイアウト数を表した片対数グラフである．

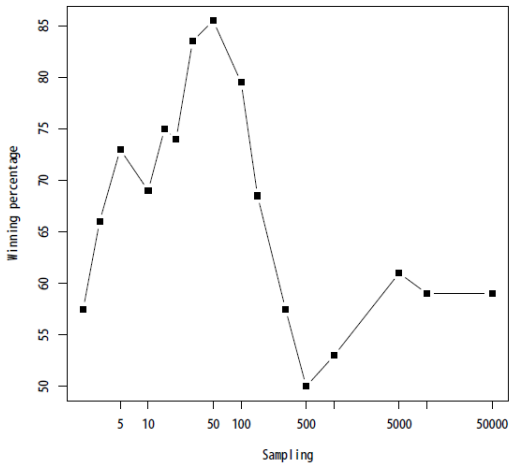

図 1：ハイブリッドモデルの勝率

5.1 シミュレーション 2: 結果

シミュレーション 2 の結果を示す．縦軸は LSVR の UCT に対する勝率を表し，横時軸は一手あたりのプレイアウト数を表した片対数グラフである．

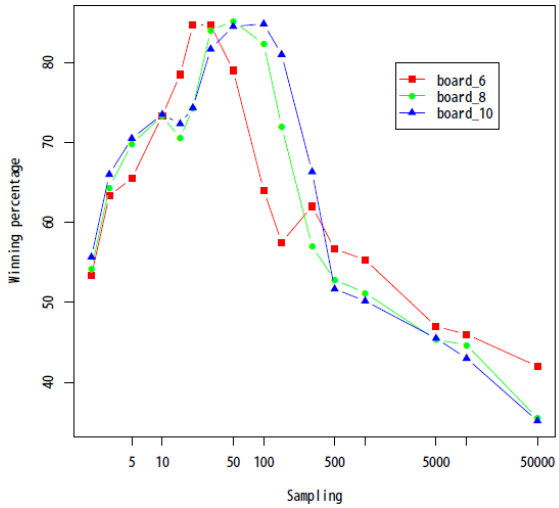

図 2: 盤面のサイズを変えた時の結果

6. 考察

シミュレーション 1 の結果から，ハイブリッドモデルにした場合 LSVR のみでモデルを構築した時に比べ，プレイアウトの回数が 500 以上の時にも勝率が 50%を下回る事は無かった．プレイアウトの回数が少ない時に強い LSVR と，多い時に強い UCT の利点を併せ持ったモデルを構築出来たと言える．盤面のサイズを変更したシミュレーション 2 では，盤面のサイズが小さい 6x6 の時に比べ，10x10 の場合勝率のピークがよりプレイアウトの多い時となっている．このことから探索空間が広く，深くなった時においても LSVR の選択肢の数が増えた場合であっても早い段階で良い手を選ぶ事が出来るという特性は発揮されており，囲碁への拡張が妥当である事が考えられる．

参考文献

1) Levente Kocsis, Csaba Szepesvari: Bandit based Monte-Carlo Planning, ECML'06 In: ECML-06, *LNCS*, 4212, pp. 282-293 (2006).
2) Peter Auer, Nicolo Casa-Bianchi, and Paul Fischer: Finite-time analysis of the multiarmed bandit problem, *Machine Learning*, 47, 235-256 (2002).
3) Sylvain Gelly, Yizao Wang, Remi Munos, Olover Teytaud,: Modification of UCT with Patterns in Monte-Carlo Go, *INRIA*, 6062 (2006).
4) Yu Kohno, Tatsuji Takahashi: Loosely Symmetric Reasoning to Cope with The Speed-Accuracy Trade-off, *The 6th International Conference on Soft Computing and Intelligent Systems, The 13th International Symposium on Advanced Intelligent Systems*.
5) 大用 庫智: ヒト認知バイアスのモンテカルロ法への応用, 2009 年度情報科学科卒業文集 (2010).
6) 篠原 修二，田口 亮，桂田 浩一，新田 恒雄: 因果性に基づく信念形成モデルと N 本腕バンディット問題への適用，人工知能学会論文誌, 22 巻 1 号, pp.58-68 (2007).
7) 西村 友伸，大用 庫智，高橋達二: 可変参照型緩対称性推論のモンテカルロ木探索での効果，ゲームプログラミングワークショップ 2012 (2012).

Copyright ©2013 Information Processing Society of Japan.
All Rights Reserved.